On the Structure of Cyclic Linear Disentangled Representations

Matthew Painter, Jonathon Hare, Adam Prügel-Bennett
Department of Electronics and Computer Science
Southampton University
{mp2u16, jsh2, apb}@ecs.soton.ac.uk

Abstract

Disentanglement has seen much work recently for its interpretable properties and the ease at which it can be induced in the latent representations of variational auto-encoders. As a concept, disentanglement has proven hard to precisely define, with many interpretations leading to different metrics which do not necessarily agree. Higgins et al. [2018] offer a precise definition of a *linear* disentangled representation which is grounded in the symmetries of the data. In this work we focus on cyclic symmetry structure. We examine how VAE posterior distributions are affected by different observations of the same problem and find that cyclic structure is encouraged even when it is not explicitly observed. We then find that better prior distributions, found via normalising flows, result in faster convergence and lower encoding costs than the standard Gaussian. We also find that linear representations can be distinguished from standard ones solely through disentanglement metrics scores, possibly due to their highly structured posteriors. Finally, we find preliminary evidence that linear disentangled representations offer better data efficiency than standard disentangled representations.

1 Introduction

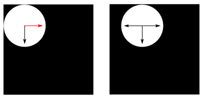
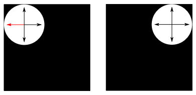
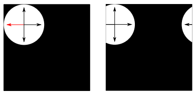
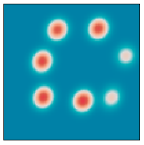
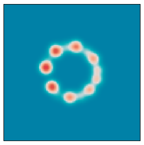
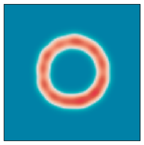
Symmetry based disentangled representation learning (SBDRL) [Higgins et al., 2018] recently provided a concrete definition for *linear* disentangled representations, a form of disentanglement that is defined with respect to symmetry groups acting on the data. The resulting representations are particularly interpretable since we can relate independent representational subspaces to the actions of symmetries in the data. These symmetries might relate to concepts such as translation of an object in the scene or its orientation. This work will focus on linear disentangled representations learnt by a variational auto-encoder (VAE) [Kingma and Welling, 2014] and their structure under the influence of cyclic symmetry structures. We first provide a brief overview of the SBDRL framework.

SBDRL VAE representation learning is concerned with the mapping from an observation space (generally images) $\mathcal{O} \subset \mathbb{R}^{n_x \times n_y}$ to a vector space $\mathcal{Z} \subset \mathbb{R}^l$ known as the latent space. SBDRL includes the additional construct of a world space $\mathcal{W} \subset \mathbb{R}^d$ containing the possible states which are represented by observations. There exists a generative process $b : \mathcal{W} \to \mathcal{O}$ and a inference process $h : \mathcal{O} \to \mathcal{Z}$, the latter being accessible and parametrised by the VAE encoder. SBDRL assumes for convenience that both h and b are injective.

SBDRL proposes to disentangle *symmetries* of the world space, the result of actions $\cdot_{\mathcal{W}} : G \times \mathcal{W} \to \mathcal{W}$ by the symmetry group $G = G_1 \times \cdots \times G_s$ on the world space. The component groups G_i reflect the individual symmetries and the particular decomposition need not be unique. SBDRL provides us with the following definition:

1st NeurIPS workshop on Interpretable Inductive Biases and Physically Structured Learning (2020), virtual.

Table 1: Definition of FlatLand variants, example samples and posteriors on an action subspace of a ForwardVAE. Arrows indicate possible actions. Red arrow indicates action chosen for that sample.

	None	Contact	Gradual
State Space	$S = \{x_i \| i \in [R, 64-R]\}$	$S = \{x_i \| i \in [R, 64-R]\}$	$S = \{x_i \| i \in [0, 64]\}$
Boundaries	$g_x \circ x_{64-R}$, $g_x^{-1} \circ x_R$ are not observed.	$g_x \circ x_{64-R} = x_R$, $g_x^{-1} \circ x_R = x_{64-R}$	$g_x \circ x_{64} = x_1$, $g_x^{-1} \circ x_1 = x_{64}$
Samples			
Posterior			

Definition 1.1. A representation space $\mathcal{Z}$ is **linearly disentangled** with respect to decomposition $G = G_1 \times \cdots \times G_s$ if:

1. There is a group action $\cdot_{\mathcal{Z}} : G \times \mathcal{Z} \to \mathcal{Z}$
2. The composition $f = h \circ b : \mathcal{W} \to \mathcal{Z}$ is equivariant with respect to the group actions on $\mathcal{W}$ and $\mathcal{Z}$. i.e. $g \cdot_{\mathcal{Z}} f(w) = f(g \cdot_{\mathcal{W}} w) \; \forall w \in \mathcal{W}, g \in G$.
3. There is a decomposition $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_s$ such that $\mathcal{Z}_i$ is fixed by the action of all $G_j, j \neq i$ and affected only by G_i

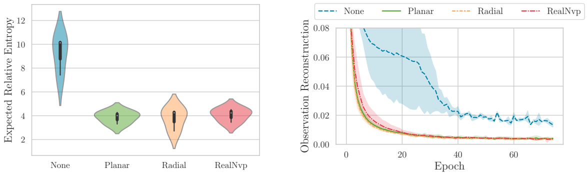
Assuming $\mathcal{Z}$ is a real vector space, SBDRL then tells us that linear disentangled representations will manifest as mappings to $\mathcal{Z}$ where actions by G_i are equivalent to *irreducible* representations. We note that irreducible representations of cyclic group C_N are rotation matrices with angle $\frac{2\pi}{N}$.

2 Cyclic Structure of Prior and Posterior

This section will look at structure of the posterior and prior distributions of linear disentangled representations with respect to group structure $G = C_N \times C_N$. Cyclic structure is the most common in disentanglement datasets, e.g. FlatLand [Caselles-Dupré et al., 2018] and dSprites [Higgins et al., 2017] which both exhibit purely cyclic symmetries.

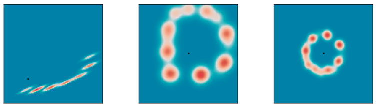
2.1 Cyclic Posterior

Prior work has considered translation of an object to be the result of acting on the world space with symmetry $C_x \times C_y$, i.e. assume the object warps to the other side when passing the image edge. We consider FlatLand, a grid world where a single white circular agent moves (with actions up, down, left or right) over a black background. We can define this problem independently (and equivalently) for C_x and C_y by a state space S, actions $\circ$ and 'contact' boundary conditions as provided in Table 1. Such structure has been learnt by ForwardVAE [Caselles-Dupré et al., 2019], a model which learns internal representations for group actions by observing transitions and optimising them to be equivariant with actions on the world space. This is a valid interpretation of the symmetry structure, we could however consider other boundary conditions such as continuous transitions or no transitions. These conditions are defined by their states S and boundary conditions in Table 1. The same table presents the posterior distribution of a ForwardVAE (4 convolutional layers + 3 linear) trained on each of the possible FlatLand variants. Whilst the model has 4 latent dimensions on this problem, we present the latent subspace related to vertical translation, allowing us to plot the posterior in 2D. We can see that all variants result in similar circular structure in the posterior. The cyclic structure of the contact and gradual warping cases is expected, however it is not immediately obvious why it is present in the no warping case since it is not explicitly present in the data.



(a) Relative entropy (b) Observation Recon

Figure 1: Comparison of priors under normalising flows.



(a) No KL (b) No flow (c) Planar

Figure 2: Posterior distribution for different flows early in training. Dot indicates origin.

We give one potential reason for this by considering expected posterior structure. Without warping, we have possible states $\{x_R, x_{R+1}, \dots, x_{N-R}\}$ and additional (local) metric knowledge that $d(x_i, x_j) \propto |i - j|$. This metric knowledge is true visually (moving by a few pixels minimally changes the image) and under the 'symmetry'[1] (moving from x_i to x_{i+3} requires acting on x_i 3 times by generator g_x but only once to reach x_{i+1}). If we assume the metric is preserved through the VAE encoding[2] then we would expect the sequence $(x_i)_i$ to be encoded in a line or curve. We see from Figure 2a that the divergence loss is required for the posterior to curve around the origin. Possibly it provides a means to minimise the description length whilst retaining representation linearity since clustering centres around the origin (marginally) raises density there. Of course, encoding Cartesian position across 4 latents instead of 2 is at odds with the minimal description length objective, however the loss pressure towards reconstructing post action latents appears to overcome this. Furthermore, the description length is determined by the prior, which we will examine next under contact boundary conditions.

2.2 Priors

Given that linear disentangled posteriors have consistent and constrained structure, which doesn't fit the form of a normal Gaussian, it seems beneficial to explore alternate priors. We will look specifically at sequentially adapting the posterior through normalising flows thus emulating a change in prior. We estimate the efficiency of different methods by comparing their expected relative entropy between posterior and prior. We will present the first 75 epochs of training, so we can compare convergence rates of flows verses the standard prior and look at how the representations evolve.

We find in Figure 1a that the expected relative entropy achieved by flow based methods is consistently lower than that under the standard Gaussian prior, showing (unsurprisingly) that a normal Gaussian is a poor choice for cyclic problems. We evidence this further in Figure 1b which reports the expected post-action observation reconstruction $||x_a - \hat{x}_a||$ for models under the different flows. We see that all flows again perform similarly whilst the standard prior performs and converges significantly worse. The major difference between posteriors for the flows and the standard prior is that the latter seems to prefer encoding along Cartesian axes, as can be seen by comparing Figure 2b to Figure 2c. To learn a strong representation under the standard prior, the model has to overcome this axis-aligning pressure at the expense of relative entropy and slower convergence.

3 Distinguishing Linear Disentangled Representations

Given the highly specific structure of the flatland linear disentangled representations, we should be able to see evidence of them in disentanglement metrics. In Figure 3 we provide the probability that an SVM with radial basis function kernels can distinguish a linear disentangled representation (one coming from ForwardVAE or RGrVAE [Painter et al., 2020]) from a standard one (VAE, β-VAE [Higgins et al., 2017], CC-VAE [Burgess et al., 2018], FactorVAE [Kim and Mnih, 2018], DIP-VAE [Kumar et al., 2017]). We find that the classical DCI disentanglement, Completeness and Informativeness metrics were the most informative classical metrics Interestingly the true independence, which measures the extent different groups act on different latent subspaces, was not a

[1] We use the term symmetry here loosely since we can not relate the sequential observations to any specific symmetry group - we have not observed actions at all states.

[2] This seem a reasonable assumption due to the variational sampling $z \sim q(z)$, at least in the visual sense.

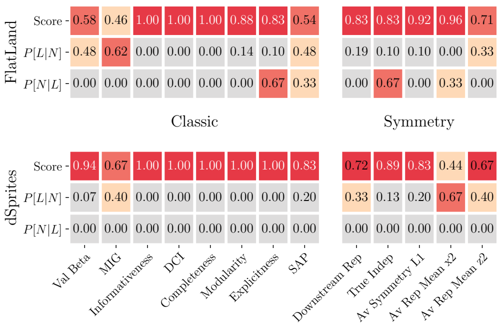
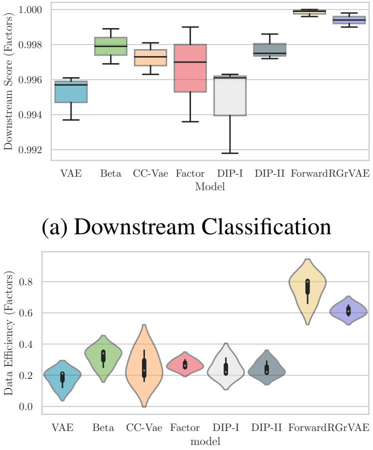


Figure 3: Relevance of each metric to distinguishing linear from classical disentanglement. Scores are the probability that an RBF-SVM can classify the model correctly from the metric alone. $P[L|N]$ is the probability the SVM predicted a model was (L)inear disentangled when it was (N)ot/(N)ormal.



(a) Downstream Classification

(b) Data Efficiency

Figure 4: Gradient boosted tree downstream task on FlatLand.

perfect discriminator on FlatLand, likely due to baselines learning good 1D representations of the actions. Another interesting result was that rotation matrix angle (trained after the fact to transform pre-action latent to post-action latent) was not a perfect discriminator (symmetry l1) as you might expect. Neither were the resulting errors in predicting the post action latent (Rep Mean z2). This may be down to large number of minima in the loss space resulting in imperfect rotation matrices being learnt for the linearly disentangled models. We did see poorer performance in the externally learnt representations compared to those learnt by the model, which supports this possibility.

4 Downstream Tasks

Locatello et al. [2018] found that disentangled representations correlated with improved performance in downstream tasks, but not with improved data efficiency, which had previously been given as a motivator for such work. The standard downstream task for disentangled representation learning (eg. Locatello et al. [2018], Caselles-Dupré et al. [2019]) is to predict generative factor values from the latent values. We will compare our models on FlatLand under this task. We should note that similar to Locatello et al. [2018], we found that the baselines did not consistently converge to disentangled representations however the majority appeared to do so.

From Figure 4a we can see that all models allowed strong downstream accuracies however the linearly disentangled models achieved consistently higher scores with lower variance than the baselines, suggesting they are stronger representations for downstream tasks. We can also see by Figure 4b that they have better data efficiency (accuracy based on 100 samples divided by that on 10000 samples) than the baselines. This suggests that the *type* of disentanglement is important to data efficiency and that the work of Locatello et al. [2018] might not capture this. We stress that this result is for a simple case and requires evaluation over more datasets and settings to be certain of this claim.

5 Conclusions

Linear disentangled representations offer much more structured distributions than standard disentanglement and using SBDRL their structure can be defined very precisely with respect to the symmetry groups acting on the data. In this work we have shown this structure explicitly for cyclic symmetries and found that the usual isotropic Gaussian is a poor choice of prior. Due to their structure we are also able to distinguish linear disentangled representations from baselines purely by their scores on disentanglement metrics. Finally we find that linear representations are better for the downstream task of predicting generative factors and offer better data efficiency on the FlatLand dataset, the latter being interesting given results that show disentanglement in classical models does not correlate with data efficiency.

References

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations, ICLR*, 2014.

Hugo Caselles-Dupré, Louis Annabi, Oksana Hagen, Michael Garcia-Ortiz, and David Filliat. Flatland: a lightweight first-person 2-d environment for reinforcement learning. *arXiv preprint arXiv:1809.00510*, 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr*, 2(5):6, 2017.

Hugo Caselles-Dupré, Michael Garcia Ortiz, and David Filliat. Symmetry-based disentangled representation learning requires interaction with environments. In *Advances in Neural Information Processing Systems*, pages 4608–4617, 2019.

Matthew Painter, Jonathon Hare, and Adam Prugel-Bennett. Linear disentangled representations and unsupervised action estimation. *arXiv preprint arXiv:2008.07922*, 2020.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.